A CUSUM to Detect Person Misfit

Tendeiro, Jorge N.; Meijer, Rob R.

Published in:
Applied Psychological Measurement

DOI:
10.1177/0146621612446305

Article

A CUSUM to Detect Person Misfit: A Discussion and Some Alternatives for Existing Procedures

Applied Psychological Measurement
36(5) 420–442

DOI: 10.1177/0146621612446305
http://apm.sagepub.com

Jorge N. Tendeiro[1] and Rob R. Meijer[1]

Abstract

This article extends the work by Armstrong and Shi on CUmulative SUM (CUSUM) person-fit methodology. The authors present new theoretical considerations concerning the use of CUSUM person-fit statistics based on likelihood ratios for the purpose of detecting cheating and random guessing by individual test takers. According to the Neyman–Pearson Lemma, the optimality of such statistics relies on how accurately normal and aberrant behaviors are modeled. General and specific models for cheating and random guessing are investigated. The detection rates of several statistics are compared using simulated data. Results showed that the likelihood-based CUSUM statistics that use the proposed models for aberrant behavior performed better than some of the more commonly used statistics, especially for cheating behavior.

Keywords

item response theory model, cheating, random guessing, aberrant behavior detection, likelihood ratio, cumulative sum

The evaluation of individual test score validity is important in education and achievement testing. For example, a frequently encountered problem is the inflation of test scores due to preknowledge of (subsets of) items. One way to check the validity of test scores is to assess the fit of an item score pattern to a test model. Item score patterns that are very unlikely are called aberrant or misfitting, and resulting test scores may not correctly reflect the ability or trait level of an examinee (Drasgow, Levine, & Williams, 1985; Meijer & Sijtsma, 2001).

The identification of misfit does not explain the source of the misfit itself. Two types of misfit that are of concern occur when examinees overperform or underperform on a subset of the items. An examinee might underperform on a cognitive test for various reasons: lack of knowledge on some of the subjects being evaluated, unfamiliarity with the language in which the test is written, or tiredness. An unexpected high score on a cognitive test may indicate that the examinee cheated successfully on some of the most difficult items or that a teacher changed incorrect item scores into correct scores (see Jacob & Levitt, 2003). Once an aberrant score

[1]University of Groningen, Netherlands

Corresponding Author:
Jorge Tendeiro, Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands
Email: j.n.tendeiro@rug.nl

pattern is identified, there is a need for inspection at the individual level to accurately pinpoint the reasons that led to the aberrant behavior. Meijer, Egberink, Emons, and Sijtsma (2008) provided an example of how to combine qualitative and quantitative information to interpret aberrant response patterns.

In the context of item response theory (IRT; Embretson & Reise, 2000) several person-fit statistics are available that are sensitive to aberrant response behavior. For an overview see, for example, Meijer and Sijtsma (2001) and Karabatsos (2003). Different statistics are sensitive to different types of aberrant response behavior. There is no method that guarantees identification of all types of aberrant item score patterns (see Meijer, 1996, and Meijer & Sijtsma, 2001, for an overview of possible types of aberrant score patterns). For example, randomly answering some of the items on a test may not result in an unlikely item score pattern. It is, therefore, important to realize that assessing person-fit is a difficult task.

In the present study, the authors focus on a group of statistics that uses statistical process control (SPC) techniques (Meijer & van Krimpen-Stoop, 2010; van Krimpen-Stoop & Meijer, 2000, 2001). These statistics are useful in identifying aberrancies which are sequential in nature. The sequence is determined by the order in which the items are given to the examinee. Taking the order of the items into consideration allows identification of subsections of the test where the examinee seems to display an unusual response behavior; hence, more detailed information is available than when considering the total number-correct score. The aim of this article is to discuss a number of SPC-based person-fit statistics, to discuss some drawbacks of these statistics, and to propose alternative ways to calculate these statistics. By means of a simulation study, the detection rates of existing and new statistics are compared.

SPC

One of the recent innovations in person-fit theory was the introduction of techniques imported from SPC (Bradlow, Weiss, & Cho, 1998; van Krimpen-Stoop & Meijer, 2001). Unlike the classical statistics which seek misfit by only focusing on some function of the *final* number-correct score, procedures relying on control charts provide information about what occurred *during* the test, on the item level. This allows researchers to identify sections of the test with unusual item score patterns. Control charts allow identification of ''local'' deviant behavior that might otherwise pass unnoticed. Consider the following situation: An examinee exhibits a score pattern with perfect scores on the first half of the test and poor scores on the second half of the test. This score pattern seems interesting enough for further investigation by the examiner (burnout?, impatience?, lack of time?). However, because the associated number-correct score is not unlikely, this pattern might not be identified by classical person-fit statistics. One of the tools used in SPC is the CUSUM (CUmulative SUM control chart; Page, 1954). A CUSUM is a sequential technique that provides information about the production process as it occurs. The main advantage is that it allows early intervention in the process once an irregular pattern is detected. Graphical control charts are especially useful to facilitate visualization of the whole control process.

CUSUM procedures were already introduced, and necessarily adapted, to IRT person-fit measurement. Bradlow et al. (1998) used a control chart methodology to identify examinees with aberrant response patterns in computerized adaptive tests (CATs). They introduced a normalized statistic which is updated after the administration of each item. After each administered item, a researcher can determine whether the statistic is unusually small or large. Upper and lower *control limits* are used to help in deciding whether a sequence of scores is to be considered aberrant. Applications of CUSUM procedures to CATs take into account that CATs are

sequential and adaptive procedures, and can therefore be regarded as ''industrial processes'' to be monitored.

van Krimpen-Stoop and Meijer (2000; see also Meijer & van Krimpen-Stoop, 2010) introduced an alternative procedure to the one proposed by Bradlow et al. (1998). The method of van Krimpen-Stoop and Meijer allowed determining upper and lower CUSUM control functions for CATs. To establish some notation, suppose that a test with n items is administered to an examinee with latent ability θ. Let X_i ($i = 1, \ldots, n$) be the Bernoulli random variable corresponding to the answer given to item i, with conditional probability function $p_i = P(X_i = 1|\theta)$. van Krimpen-Stoop and Meijer (2000) defined the iterative ''upper'' and ''lower'' cumulative statistics as

$$C_i^+ = \max\{0, T_i + C_{i-1}^+\}, \tag{1}$$

$$C_i^- = \min\{0, T_i + C_{i-1}^-\}, \tag{2}$$

for $i = 1, \ldots, n$ and $C_0^+ = C_0^- = 0$. $T_i = \frac{1}{n}\frac{X_i - p_i}{\lambda_i}$ is a function of weighted differences between observed and expected item scores at stage i, corrected for test length. For example, the weights λ_i may equal the estimated standard deviation of the residuals or may equal the square root of the test information function. T_i can be evaluated at the updated ability estimate $\hat{\theta}_{i-1}$ or alternatively can be evaluated at the final ability estimate $\hat{\theta}_n$. The CUSUM procedure by van Krimpen-Stoop and Meijer (2000) also required estimating upper and lower control limits, $U_i(\alpha)$ and $L_i(\alpha)$, respectively. A sequence of scores is classified as aberrant if, at any step i, $C_i^- \leq L_i(\alpha)$ or $C_i^+ \geq U_i(\alpha)$.

Likelihoods for Normal and Aberrant Models

A different way of estimating T_i takes aberrant behavior into account (Drasgow, Levine, & Zickar, 1996). Assume that the observed scores on a set of items are independent, conditional on the ability parameter; this is the so-called *local independence* assumption. It is assumed that local independence holds for normal and aberrant behavior. Let $L_{\text{normal}}(\theta|\boldsymbol{x}, \boldsymbol{\phi}) = \prod_i p_i^{x_i}(1 - p_i)^{1-x_i}$ denote the likelihood of a normal response vector $\boldsymbol{x} = (x_1, \ldots, x_n)$; $\boldsymbol{\phi}$ denotes the vector of all item parameters. Assume that the probability of correctly answering each item i is modeled under aberrant behavior, which is denoted as p_i^*. The likelihood of an aberrant response pattern is $L_{\text{aberrant}}(\theta|\boldsymbol{x}, \boldsymbol{\phi}) = \prod_i (p_i^*)^{x_i}(1 - p_i^*)^{1-x_i}$. The Neyman–Pearson Lemma (Neyman & Pearson, 1933) states that the likelihood ratio

$$\frac{L_{\text{aberrant}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})}{L_{\text{normal}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})} \tag{3}$$

provides an optimal statistic to test normal versus aberrant behavior. The test is optimal in the sense that it maximizes the power to detect aberrance for fixed Type I error. This optimality is only valid if normal and aberrant behaviors are accurately modeled, and if the (fixed) final ability estimate $\hat{\theta}_n$ is used. The Neyman–Pearson Lemma will hold only locally if ability estimates are updated after each item administration.

In cases where it can be assumed that the examinees have been sampled from a distribution $F(\theta)$, θ can be integrated out from $L_{\text{aberrant}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})$ and $L_{\text{normal}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})$ (Drasgow et al., 1996). Equation 3 can be rewritten as a ratio of *marginal* likelihoods:

$$\frac{\int_\Theta L_{\text{aberrant}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})\, dF(\theta)}{\int_\Theta L_{\text{normal}}(\theta|\boldsymbol{x}, \boldsymbol{\phi})\, dF(\theta)}. \tag{4}$$

An Application of the Likelihood Ratio

Armstrong and Shi (2009) proposed a model where CUSUM statistics similar to Equations 1 and 2 were used, except that the updates T_i result from logarithms of likelihood ratios derived from Equation 3. For aberrant behaviors where overperformance of some kind is sought, it can be seen that

$$\gamma_i^U = \ln \frac{(p_i^U)^{x_i}(1-p_i^U)^{1-x_i}}{p_i^{x_i}(1-p_i)^{1-x_i}}, \tag{5}$$

$$C_i^U = \max\{0, \gamma_i^U + C_{i-1}^U\}.$$

Probability p_i^U denotes the probability of a correct response for the aberrant overperformance profile under investigation. Notice that γ_i^U is more likely to be positive in case of aberrant behavior, whereas in case of normal response behavior, it is more likely that γ_i^U attains negative values. The formula for C_i^U does not allow negative values. Hence, C_i^U detects overperformances by increasing its score accordingly. If the sum of accumulated deviances crosses an upper threshold (to be estimated), then the score pattern will be flagged as aberrant.

In cases where the aberrant behavior is described as an underperformance of the examinee, the lower CUSUM statistic equals

$$\gamma_i^L = \ln \frac{p_i^{x_i}(1-p_i)^{1-x_i}}{(p_i^L)^{x_i}(1-p_i^L)^{1-x_i}}, \tag{6}$$

$$C_i^L = \min\{0, \gamma_i^L + C_{i-1}^L\}.$$

Probability p_i^L denotes the probability of a correct response for the aberrant underperformance of interest. Note that the γ_i^L ratio is defined as the inverse of the general expression in Equation 3 so that aberrancies are reflected by negative values of γ_i^L (similarly to the C_i^- statistic). Hence, γ_i^L is more likely to be negative in cases of aberrant behavior. C_i^L accumulates all deviances of the score pattern from what is expected due to underperformance of the examinee. If C_i^L crosses the lower CUSUM limit (to be estimated), then the score pattern is flagged as aberrant.

A two-sided control statistic which combined C_i^U and C_i^L was also proposed by Armstrong and Shi (2009):

$$C_{\max}^U = \max\{C_i^U\} \quad \text{and} \quad C_{\min}^L = \min\{C_i^L\}, \quad i = 1, \ldots, n,$$

$$C^{LR} = C_{\max}^U - C_{\min}^L. \tag{7}$$

A score pattern is out of control whenever C^{LR} is larger than the upper CUSUM limit.

Modeling p_i and p_i^*

The optimality of the likelihood ratio statistic to detect aberrant patterns is dependent on how accurately the probabilities of a correct response under normal and aberrant behaviors are modeled. Choice of the model is therefore important. This choice may be guided by well-established traditions in the field, by empirical findings, or by a combination of both types of arguments. Some possibilities for normal and aberrant behaviors are discussed in this section. Special

attention is paid to the quadratic model for aberrant behavior, which was introduced by Armstrong and Shi (2009).

Modeling p_i

There are well-established models in IRT for dichotomous normal behavior scores, such as the three-parameter logistic (3PL) model (Birnbaum, 1968),

$$p_i = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}, \tag{8}$$

where a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the asymptotic probability of a correct response for arbitrarily small θ (c_i is also known as the ''guessing'' parameter). Other often-used models are the two-parameter logistic (2PL) model (Equation 8 with $c_i = 0$) and the one-parameter logistic (1PL) or Rasch model (Equation 8 with $c_i = 0$, $a_i = 1$), see Embretson and Reise (2000).

Modeling p_i^*

Modeling misfit in IRT is not straightforward. Each type of aberrant behavior may have its own characteristics. Two possible approaches that are available in the literature are discussed. One approach presents a general formula that may adapt, after adjustment of parameters, to most of the aberrant behaviors of interest; the other approach tries to model aberrancies individually.

Quadratic Modeling

Armstrong and Shi (2009) avoided specifying p_i^* for each type of aberrant behavior. They proposed to model p_i^U and p_i^L as quadratic functions of p_i (Figures 1 and 2):

$$g_i^U(p_i) = r_i^U p_i^2 + s_i^U p_i + t_i^U \tag{9}$$

$$g_i^L(p_i) = r_i^L p_i^2 + s_i^L p_i + t_i^L. \tag{10}$$

Probability p_i is assumed to follow the 3PL model. The support for both functions is the interval $[c_i, 1]$. Function g_i^U must satisfy (a) $g_i^U(c_i) \geq c_i$, (b) $g_i^U(p_i) > p_i$ for $c_i < p_i < 1$, and (c) $g_i^U(1) = 1$, whereas g_i^L must satisfy (a) $g_i^L(c_i) = c_i$, (b) $g_i^L(p_i) < p_i$ for $c_i < p_i < 1$, and (c) $g_i^L(1) \leq 1$. Functions g_i^U and g_i^L were considered flexible enough to model most types of aberrancies, after an appropriate estimation of parameters r_i, s_i, and t_i. However, the authors recognized that there is no rationale supporting this model for aberrant shifts.

Armstrong and Shi (2009) presented an algorithm for estimating parameters r_i, s_i, and t_i. Note that three points are required to completely identify an upper quadratic function $g_i^U(p_i)$ (similarly for $g_i^L(p_i)$). Hence, the challenge is to find three suitable points. For estimating $g_i^U(p_i)$, for instance, the three points used were $P_1 = (1, 1)$; $P_2 = (p_i(\theta_i^{\max}), p_i(\theta_i^*))$; and $P_3 = (c_i, (1 - v'')c_i + v''(p_i(\theta_i^*) + \lambda_i(c_i - p_i(\theta_i^{\max}))))$, where $\theta_i^{\max}$ is the value for θ_i which maximizes the information function $I_i(\theta)$ for item i, $\theta_i^* = \theta_i^{\max} + v'/\sqrt{I_i(\theta_i^{\max})}$, and $\lambda_i = [1 - p_i(\theta_i^*)]/[1 - p_i(\theta_i^{\max})]$. Figure 3 (based on Figure 3 in Armstrong & Shi, 2009, p. 399) illustrates function $g_i^U(p_i)$. Parameter $v' \geq 0$ determines the vertical position of P_2 above the 45° line L: P_2 is on line L when $v' = 0$, and it approaches the horizontal line $p_i^*(\theta_i) = 1$ when v'

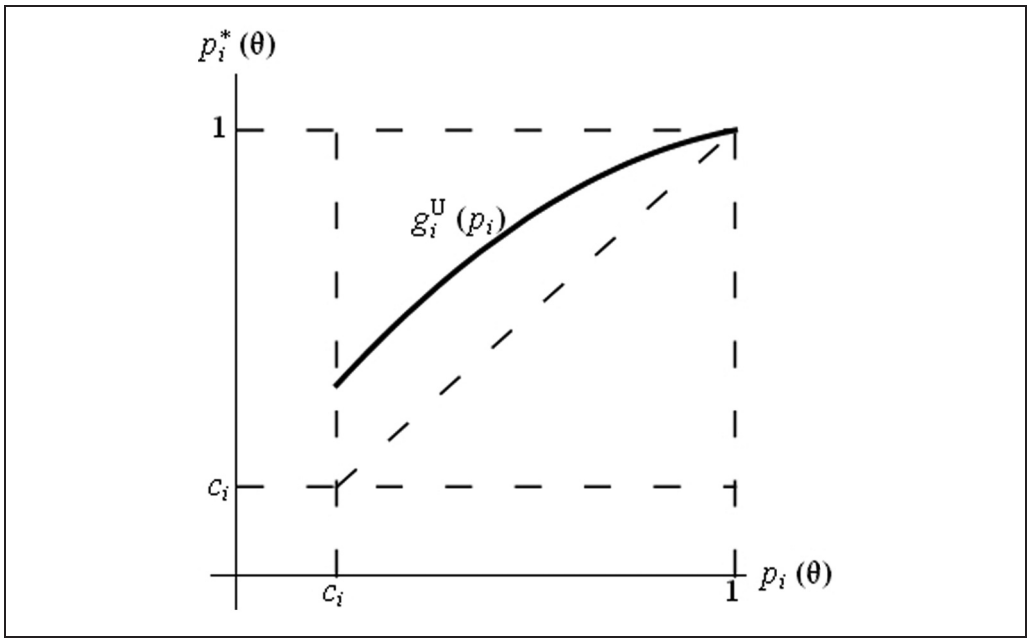

Figure 1. Quadratic upper aberrant curve

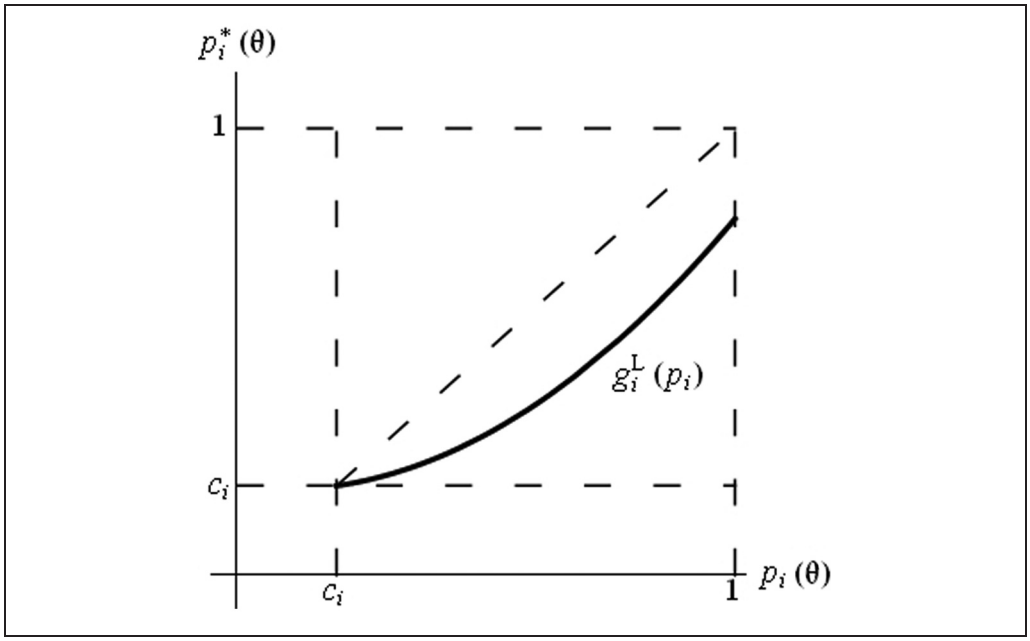

Figure 2. Quadratic lower aberrant curve

approaches infinity. Parameter λ_i is the slope of the line connecting P_1 and P_2. Parameter v'' in the interval $[0, 1]$ determines the vertical position of P_3 between the ordinates c_i ($v'' = 0$) and Q ($v'' = 1$). Function $g_i^L(p_i)$ can be estimated in a similar fashion as presented for $g_i^U(p_i)$.

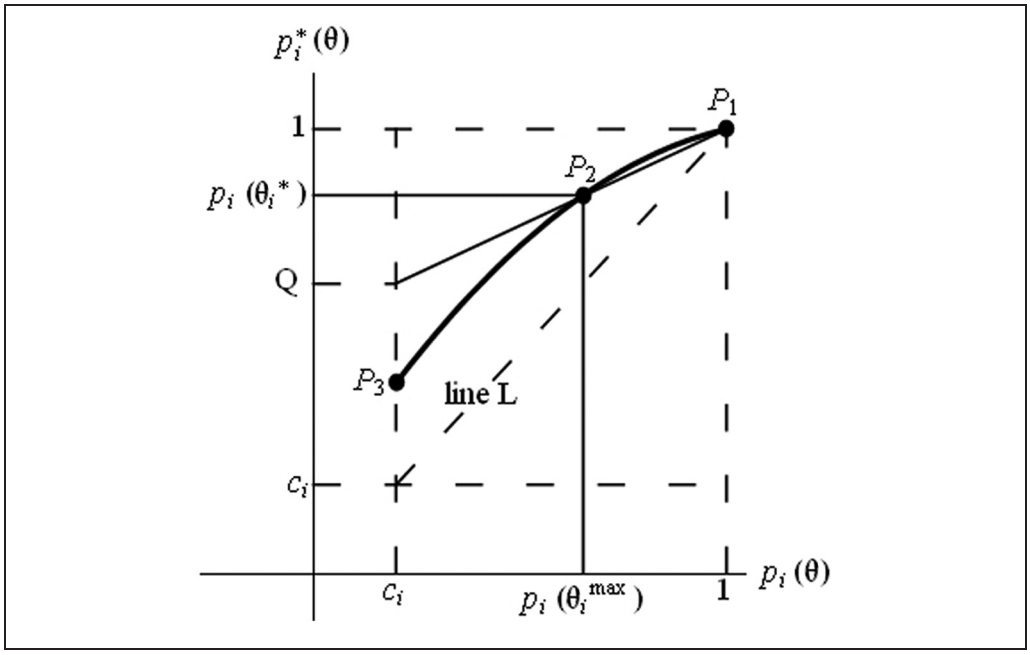

Figure 3. Quadratic model

Although simulation results showed that this approach performed well to detect aberrant response behavior (Armstrong & Shi, 2009), it may sometimes lead to problematic modeling of item responses as discussed below.

Empirical Justification

Armstrong and Shi (2009) did not present a rationale for using Equations 9 and 10 to model p_i^*. Given an item with parameters a_i, b_i, and c_i (under the 3PL model for normal behavior), the models for overperformance (Equation 9) and for underperformance (Equation 10) are fully identified. A drawback may be that *any* type of overperformance results in the same estimated model, and similarly any type of underperformance results in the same estimated model.

Consider the following specific item i as an illustration: $a_i = 1$, $b_i = 1.5$, $c_i = .2$. The 3PL model for p_i is pictured in Figure 4. The estimated upper quadratic model using $v' = v'' = .5$ is $p_i^* = g_i^U(p_i) = -.63p_i^2 + 1.47p_i + .16$. The plot of p_i^* against θ is shown in Figure 5. It is observed that the quadratic model estimates any overperformance by the same function, plotted in Figure 5. The lower asymptote is approximately .429, which does not have a natural interpretation. Overall, it seems difficult to justify the adoption of the quadratic model merely based on its mathematical formulation. This problem can be disregarded if the detection power associated with this model is satisfactory. Addressing this question is one of the goals of the simulation study presented in this article.

Flexibility

The estimation algorithm for the quadratic model is insensitive to changes in the discrimination and difficulty parameters a_i and b_i, respectively. This means that the estimation is invariant for

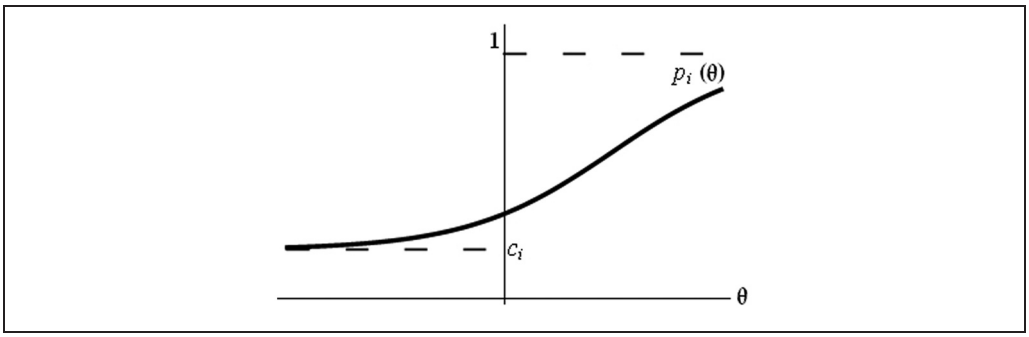

Figure 4. Three-parameter logistic (3PL) model of correct response; $a_i = 1$, $b_i = 1.5$, $c_i = .2$

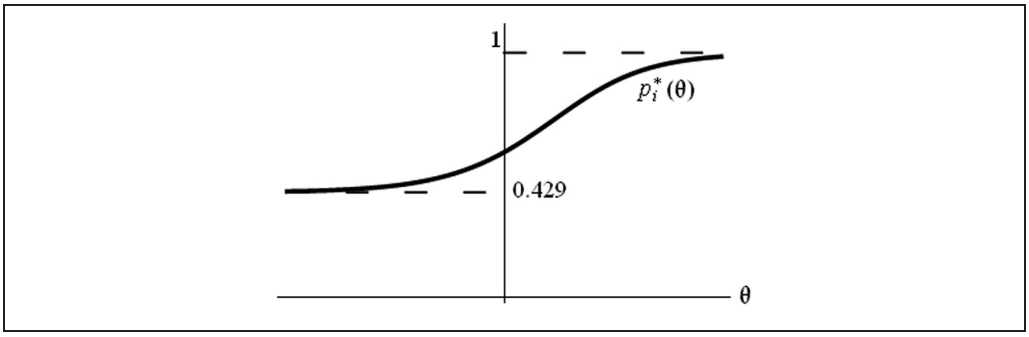

Figure 5. Aberrant model of correct response; $a_i = 1$, $b_i = 1.5$, $c_i = .2$, $v' = v'' = .5$

changes in a_i and b_i, as long as v', v'', and c_i are fixed. A mathematical proof is given in Appendix A. This property is referred to as the ''invariance'' property of the quadratic model.

One consequence of the invariance property is that the algorithm suggested by Armstrong and Shi (2009) to estimate Equation 9 always leads to the same model of aberrant behavior given the values for the guessing parameter c_i and the constants v', v''. A similar result applies to estimating the quadratic form in Equation 10 for underperformance aberrant behavior. It is questionable whether this is realistic. For example, an examinee cheating on a cognitive test may, as a result of preknowledge, cheat on the more difficult items or on the items of moderate difficulty (Jacob & Levitt, 2003). The estimation method for the quadratic model results in the same model for two or more aberrant behaviors, even though these behaviors might be quite different in nature.

Mathematical Boundaries of Equations 9 and 10

The possibility that function $g_i^U(p_i)$ (respectively $g_i^L(p_i)$) is larger than 1 (respectively smaller than c_i) is not mathematically excluded. For example, when $a_i = 3$, $b_i = 2$, $c_i = 0.33$, $v' = 0.9$, and $v'' = 0.5$, it can be seen that $g_i^U(p_i) > 1$ for $\theta > 2.413$ and that $g_i^L(p_i) < c_i$ for $\theta < 1.878$ ($g_i^L(p_i)$ is plotted in Figure 6).

Such models have no sensible interpretation, hence, they are poor models for probabilities of correct response under (upper and lower) aberrancies. Although it can be argued that a different choice of parameters v' and v'' could solve this problem, the estimation method should be

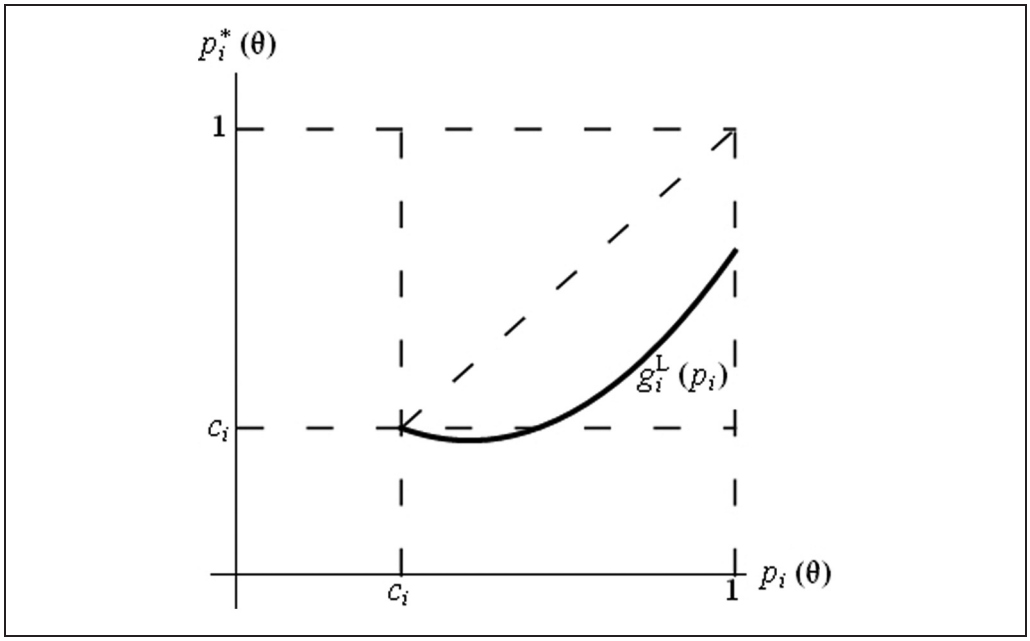

Figure 6. Quadratic lower aberrant curve: $a_i = 3$, $b_i = 2$, $c_i = .33$, $v' = .9$, $v'' = .5$

robust enough to avoid this kind of anomaly. This problem is referred to as the ''boundary'' problem associated with the quadratic model.

It is possible to verify whether the estimated quadratic function is adequate after estimating the parameters r_i, s_i, and t_i. Rewriting Equation 9 as

$$g_i^U(p_i) = r_i^U \left[p_i - \left(-\frac{s_i^U}{2r_i^U} \right) \right]^2 + \left(t_i^U - \frac{(s_i^U)^2}{4r_i^U} \right), \tag{11}$$

and using the property that $g_i^U(p_i)$ is a monotonic increasing function in the interval $[c_i, 1]$, it can be verified that a necessary and sufficient condition for $g_i^U(p_i) \leq 1$ is $-s_i^U/2r_i^U \geq 1$ or, equivalently,

$$s_i^U + 2r_i^U \geq 0. \tag{12}$$

Similarly, a necessary and sufficient condition for $g_i^L(p_i) \geq c_i$ is $-s_i^L/2r_i^L \leq c_i$ or, equivalently,

$$s_i^L + 2r_i^L c_i \geq 0. \tag{13}$$

In case this condition is not met, a different choice of v' and/or v'' should be made. In general, increasing the value of v'' solves the problem, as increasing v'' has the effect of ''straightening'' the quadratic curve. In the extreme situation $v'' = 1$, the curve is a straight line and the boundary problem is no longer an issue.

Case-by-Case Modeling

A less ambitious endeavor than fitting a global model to a large family of types of aberrant behavior is to model each type of aberrant behavior individually. There are advantages as well

as drawbacks inherent in this approach. A clear disadvantage is that one should define different models for each type of misfit. As a consequence, several person-fit statistics might be needed as different statistics may be optimal for detecting different aberrancies. Adjusting the alpha level of each procedure to maintain the desired false-positive rate might be needed. However, statistics which are specifically built for detecting a special type of aberrant behavior have, in general, higher probability of detecting such cases than other statistics.

For example, Drasgow et al. (1996; see also Levine & Drasgow, 1988) modeled specific types of aberrant responding such as cheating, dissimulation (e.g., on personality tests), unfamiliarity with computerized tests, or randomly answering some of the items. Models for other types of aberrant behavior can be considered, for example, sleeping behavior, alignment errors, or plodding behavior (Meijer, 1996), test anxiety (Green, 2011; Rulison & Loken, 2009), or cheating from a neighbor (Belov, 2011).

Importance of Modeling p_i^* in Log-Likelihood Ratio CUSUM Statistics

One must be extremely careful when modeling normal and aberrant behaviors. The optimality of the Neyman–Pearson Lemma in the context of likelihood ratios strongly depends on the accuracy of both models. Inaccurate models may result in increased Type I or Type II errors, which is an undesirable outcome in person-fit measurement.

A simulation study was carried out to assess how much the detection rate of CUSUM procedures using logarithms of likelihood ratios can be influenced by alternative modeling of p_i^*. Specifically, the authors wanted to assess whether the detection power of log-likelihood ratio CUSUM statistics was overly affected by replacing the (general) quadratic model for p_i^* with alternative models that are tailored for specific types of aberrant behavior. They focused on two types of aberrant behavior due to their relevance in education and achievement testing: cheating and random guessing. Models for the probability of a correct response under each of these aberrancies are introduced in the next section. As argued, these alternative models are based on simple empirical reasoning. Hence, practical advantage can be achieved in case these models for cheating and random guessing do behave better than, or at least similar to, the quadratic model.

Simulation Study

The authors present models for the probability of a correct response under random guessing and cheating. These models were used as alternatives to the quadratic model, with the objective of checking how the detection rate of aberrant behavior is affected by different models for p_i^* in the setting of log-likelihood ratio CUSUM statistics.

Alternative Models for p_i^*: Random Guessing and Cheating

Suppose that item i is answered randomly due, for example, to lack of knowledge or lack of time. It is reasonable to assume that item i is answered correctly with *constant* positive probability. Under the 3PL model such a probability is, precisely, the guessing parameter c_i. The value of c_i may be equal to $1/m_i$, where m_i is the number of response alternatives; larger than $1/m_i$ (when one of the alternatives can be clearly discarded); or smaller than $1/m_i$ (when one of the wrong alternatives is particularly attractive). Assume that p_i^* is equal to c_i for all ability levels:

$$p_i^*(\theta) = c_i \text{ for all } \theta. \tag{14}$$

It should be noted that, although Equation 10 seems adequate for fitting this function, the estimation procedure of Armstrong and Shi (2009) will not result in such a form because it would require $v' = \infty$.

Another important type of aberrant behavior that is often mentioned in the literature is cheating (e.g., Belov, 2011; Drasgow et al., 1996; Meijer, 1996). Typically, cheating allows an examinee to perform above his or her true ability. This may be due to preknowledge of items prior to the test or poor surveillance during the test. When an examinee cheats while answering an item, it is reasonable to assume that the discrimination and the difficulty of the item, as well as the ability of the examinee, play a minor role in predicting the probability of correct response. When the correct answer is known, this probability equals 1: $p_i^*(\theta) = 1$ for all ability levels θ. When it is assumed that an examinee's memory is not perfect, then the model can be

$$p_i^*(\theta) = d_i \text{ for all } \theta, \tag{15}$$

where the constant d_i can be chosen close to 1. It can be observed that Equation 9 is not a suitable model for this function, because condition (b) ($g_i^U(p) > p$ for $c_i < p < 1$) is being violated.

Both models in Equations 14 and 15 were used in the simulation study as alternatives to the quadratic model.

Generation Procedure

Item difficulties and ability parameters were randomly drawn from the standard normal distribution. Item discrimination and guessing parameters were randomly drawn from uniform distributions in the intervals $(.5, 1.5)$ and $(0, .25)$, respectively. These distributions and values were chosen because they cover the most usual ranges in real IRT practice (see, for example, Embretson & Reise, 2000). The number of items and examinees generated was $n = 100$ and $N = 10,000$, respectively. An $N \times n$ data set (henceforth denoted DSet) was generated using the 3PL model. Also, a calibration data set was independently generated; this data set was used to estimate the control limits associated with each person-fit statistic considered in this study. To build the calibration set, the authors simulated 100,000 response vectors for abilities randomly drawn from the standard normal distribution and using the item parameters described before. The size of the calibration set was considered large enough to provide accurate estimates of the control limits for each of the person-fit statistics to be used in the simulation study.

Statistics

Besides van Krimpen-Stoop and Meijer's (2000) and Armstrong and Shi's (2009) CUSUM statistics (C^+, C^-, C^U, C^L, C^{LR}), the authors used two log-likelihood ratio CUSUM statistics specifically designed for random guessing and cheating aberrant behaviors (C^{RR}, C^{Ch}). The difference between these statistics and those statistics in Armstrong and Shi was the model for p_i^* (Equations 14 and 15). As explained before, they wanted to verify whether a different, intuitively simpler way, of modeling p_i^* as c_i (in case of random guessing) or d_i (in case of cheating) would affect the detection rates of aberrant response behavior. A new CUSUM person-fit statistic, C_{VM}^{LR}, was also considered. This two-sided statistic is similar to the C^{LR} statistic (see Equation 7), but uses the upper and lower CUSUM statistics proposed by van Krimpen-Stoop and Meijer:

$$C_{\max}^+ = \max\{C_i^+\} \quad \text{and} \quad C_{\min}^- = \min\{C_i^-\}, \quad i = 1, \ldots, n, \tag{16}$$

$$C_{VM}^{LR} = C_{\max}^+ - C_{\min}^-. \tag{17}$$

Finally, the authors used three traditional person-fit statistics (not CUSUM based). Statistics U (Wright & Stone, 1979) and W (Wright, 1980) are residual based and are defined as

$$U = \sum_{i=1}^{n} \frac{(X_i - p_i)^2}{np_i(1 - p_i)} \tag{18}$$

and

$$W = \frac{\sum_{i=1}^{n} (X_i - p_i)^2}{\sum_{i=1}^{n} p_i(1 - p_i)}. \tag{19}$$

The likelihood-based statistic l_z (Drasgow et al., 1985) was also used in this study, and it is defined as

$$l_z = \frac{l_0 - E(l_0)}{[\mathrm{Var}(l_0)]^{1/2}}, \tag{20}$$

where $l_0 = \sum_{i=1}^{n} [X_i \ln(p_i) + (1 - X_i) \ln(1 - p_i)]$ and $E(l_0)$ and $\mathrm{Var}(l_0)$ are the expectation and variance of l_0, respectively:

$$E(l_0) = \sum_{i=1}^{n} p_i \ln(p_i) + (1 - p_i) \ln(1 - p_i) \tag{21}$$

and

$$\mathrm{Var}(l_0) = \sum_{i=1}^{n} p_i(1 - p_i) \left[\ln\left(\frac{p_i}{1 - p_i} \right) \right]^2. \tag{22}$$

More details concerning these statistics can be found, for example, in Meijer and Sijtsma (2001) and Karabatsos (2003).

The item and ability parameters of DSet were calibrated with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). The authors used the default options of BILOG-MG with two exceptions: NPArm was set to 3 to specify the 3PL model, and NALt was set to 5 to reflect their assumption that each item has 5 alternative answer options. The estimated item parameters from DSet were used to estimate the ability parameters from the calibration data set. This ensured that the scores from the simulated examinees for calibration corresponded to the same items as the scores from DSet. Next, the appropriate (upper or lower) 1% and 5% control limits for each person-fit statistic in this study were estimated: C^+, C^-, C_{VM}^{LR}, C^U, C^L, C^{LR}, C^{RR}, C^{Ch}, U, W, and l_z. This was done by computing each person-fit statistic for each simulated examinee in the calibration set and then taking the adequate 1% and 5% quantiles from the empirical distributions. For the statistics which required a model for p_i^* (C^U, C^L, C^{LR}, C^{RR}, and C^{Ch}), the authors proceeded as follows. Armstrong and Shi's (2009) quadratic model for p_i^* was estimated using the procedure described previously with $v' = .50$, $v'' = .75$. The latter values (especially v'') were used to avoid the boundary problems $p_i^* > 1$ or $p_i^* < c_i$. Necessary and sufficient conditions (Equations 12 and 13) were used to ascertain that no boundary violation occurred. The probability of a correct response for random guessing was defined by $p_i^* = c_i$, and the probability of a correct response for cheating was defined by $p_i^* = d$, where d was sampled from the uniform distribution in the interval (.91, .99). Probability d was kept fixed per examinee and across items. The interval (.91, .99) was tentatively chosen so that the probability of correct response under

cheating would be high but not exactly equal to 1, to mimic the situation where examinees can still make mistakes while cheating. At the same time, this decision avoids the mathematical impossibility of defining log-likelihood ratios when either the numerator or the denominator is exactly zero (see Equations 5 and 6).

Aberrant Response Behavior

Two types of aberrant behavior were simulated: random guessing and cheating. Each aberrant behavior was treated separately, to better characterize the performance of each statistic under each aberrant behavior. Hence, independent data sets were generated for each situation. Person-fit statistics C^{-}, C_{VM}^{LR}, C^{L}, C^{LR}, C^{RR}, U, W, and l_z were used for detecting random guessing; statistics C^{+}, C_{VM}^{LR}, C^{U}, C^{LR}, C^{Ch}, U, W, and l_z were used for detecting cheating.

All examinees selected to display aberrant response behavior had estimated ability below 0.5, as lower ability examinees are typically more prone to engage in either cheating or random guessing. The percentage of examinees whose response vectors were changed was set at three levels: SubPC = 1%, 5%, and 10% of the $N = 10{,}000$ examinees where SubPC is the percentage of examinees who had their response vector altered. These examinees were randomly chosen among those with moderately low ability estimates (<0.5). Three different proportions of changed item scores were considered: AnsPC = 5%, 10%, and 25% of the $n = 100$ items, where AnsPC is the percentage of items that were altered. These items constituted random sequences within the n-length response vector. For example, when AnsPC = .10, the program randomly picked one of the following sequences of 10 consecutive items: 1-10, 11-20, 21-30, . . . , 91-100. The 10 items chosen were altered according to the specific aberrant behavior under study. The same proportion AnsPC of aberrant scores was used within each data set.

The item and ability parameters were reestimated after the aberrant behavior imputation was completed; these estimates were used in the computations of all person-fit statistics. The correlation between parameter estimates from before and after the imputation of aberrant behavior was computed to control whether aberrance imputation overly affected parameter estimation.

Summarizing, the authors used a 2 (aberrant behavior under study) × 3 (proportion of aberrant response vectors) × 3 (proportion of changed item scores) completely crossed design. Ten replications per cell were used, hence a total of 180 data sets were analyzed. Note that DSet was the basic data set from which all the 180 data sets were derived. All the programs used in this simulation study were written in R (R Development Core Team, 2009).

Simulation Results

The item parameters were estimated for each simulated data set after the imputation of aberrant behavior. It was expected that these estimates would be very close from the estimates prior to inputting aberrancies. The correlations between the estimates of item parameters before and after inputting aberrancies were typically larger than .98. The exception was the c parameters when SubPC = 10% and AnsPC = 25% (the correlations lowered to values close to .95). Overall, these values show that the imputation of aberrant scores was kept at controlled levels. Also, the correlations between ability estimates before and after generating aberrant behavior were usually larger than .98. Thus, the imputation of aberrant behavior did not change the original score structure as shown in DSet.

Detection of aberrant behavior was done using 1% and 5% control limits for each statistic. The rate of false positives (identifying ‘‘normal’’ examinees as ‘‘aberrant’’) is given in Table 1. Typically, these values fluctuated around 1% and 5%, as expected. However, for statistics U, W, and l_z, the rate of false positives seemed to decrease with the increase of SubPC and

Table 1. Percentage Rate of False Positives for Random Guessing and for Cheating

Random guessing			Cheating		
	1%	5%		1%	5%
C^{-}	0.92 (0.04)	4.50 (0.07)	$C+$	0.94 (0.03)	4.96 (0.20)
C_{VM}^{LR}	0.86 (0.04)	4.79 (0.08)	C_{VM}^{LR}	0.89 (0.03)	4.84 (0.13)
C^{L}	0.77 (0.03)	4.60 (0.09)	C^{U}	0.82 (0.06)	5.06 (0.18)
C^{LR}	0.87 (0.02)	4.58 (0.07)	C^{LR}	0.89 (0.05)	4.60 (0.12)
C^{RR}	0.92 (0.03)	4.73 (0.10)	C^{Ch}	0.96 (0.05)	5.15 (0.16)
U	0.87 (0.04)	5.30 (0.10)	U	0.82 (0.08)	5.05 (0.30)
W	0.97 (0.06)	5.01 (0.11)	W	0.92 (0.10)	4.80 (0.27)
l_z	1.00 (0.06)	5.16 (0.12)	l_z	0.95 (0.10)	4.97 (0.31)

Note: SubPC = percentage of examinees who had their response vector altered; AnsPC = percentage of items that were altered. The values in each cell are the mean and standard deviation of 90 rates of false positives (across the three levels of SubPC, the three levels of AnsPC, and the 10 replications).

Table 2. Main Effects of SubPC and AnsPC on Detection Rates (Using 5% Control Limits), for Random Guessing and Cheating

Random guessing			Cheating		
	SubPC	AnsPC		SubPC	AnsPC
C^{-}	$F = 0.23$	$F = 909.72^{**}$	$C+$	$F = 2.41$	$F = 4371.70^{**}$
	$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .95$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$
C_{VM}^{LR}	$F = 3.31^{*}$	$F = 2585.09^{**}$	C_{VM}^{LR}	$F = 0.89$	$F = 3211.60^{**}$
	$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .98$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$
C^{L}	$F = 0.69$	$F = 6155.07^{**}$	C^{U}	$F = 5.83^{**}$	$F = 15346.40^{**}$
	$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = 1.00$
C^{LR}	$F = 1.06$	$F = 4054.01^{**}$	C^{LR}	$F = 0.43$	$F = 7864.77^{**}$
	$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$
C^{RR}	$F = 0.45$	$F = 5725.10^{**}$	C^{Ch}	$F = 1.88$	$F = 6219.23^{**}$
	$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .93$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .99$
U	$F = 4.52^{*}$	$F = 333.79^{**}$	U	$F = 2.36$	$F = 125.57^{**}$
	$\hat{\omega}^2 = .01$	$\hat{\omega}^2 = .87$		$\hat{\omega}^2 = .01$	$\hat{\omega}^2 = .73$
W	$F = 3.41^{*}$	$F = 259.22^{**}$	W	$F = 0.94$	$F = 139.73^{**}$
	$\hat{\omega}^2 = .01$	$\hat{\omega}^2 = .84$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .76$
l_z	$F = 3.46^{*}$	$F = 303.20^{**}$	l_z	$F = 1.40$	$F = 132.52^{**}$
	$\hat{\omega}^2 = .01$	$\hat{\omega}^2 = .86$		$\hat{\omega}^2 = .00$	$\hat{\omega}^2 = .74$

Note: SubPC = percentage of examinees who had their response vector altered; AnsPC = percentage of items that were altered. F values have associated $df = 2, 85$.
*significant at $\alpha = .05$. ** significant at $\alpha = .01$.

AnsPC. Thus, these statistics were more sensitive to the gradual change of the scores as aberrant behavior was inputted.

The authors investigated whether the proportion of aberrant response vectors in the data set (factor SubPC) and the length of the sequence of aberrant item scores (factor AnsPC) had an effect on the detection rates, for $\alpha = .05$. They started by fitting 3×3 full factorial models (main effects SubPC and AnsPC, and interaction effect SubPC $\times$AnsPC) for each person-fit statistic, under random guessing and cheating. In all cases, there was no significant interaction effect ($\hat{\omega}^2 = 0$). They therefore reestimated the ANOVA models considering only the main effects of SubPC and AnsPC on the detection rates. The results are summarized in Table 2.

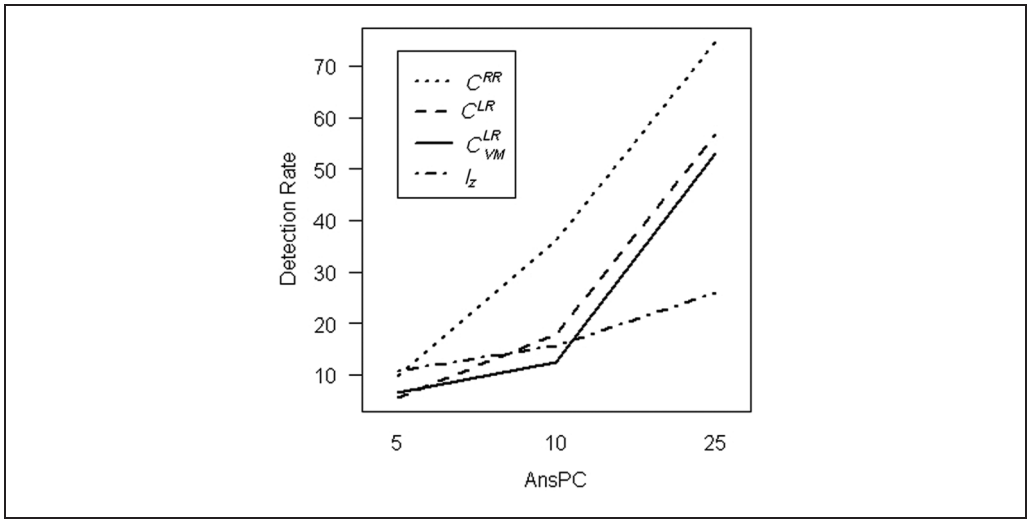

Figure 7. Random guessing: Effect of the length of the sequence of aberrant item scores on detection rates for four person-fit statistics ($\alpha = .05$, SubPC = 1%)
Note: SubPC = percentage of examinees who had their response vector altered.

It can be verified that SubPC had no effect on the detection rate of CUSUM-based statistics except for statistic C^{LR}_{VM} under random guessing. SubPC did have an effect on the detection rates of statistics U, W, and l_z under random guessing. More specifically, increasing the proportion of aberrant response scores resulted in a *decrease* of detection rates for U, W, and l_z. The effect size of SubPC on the detection rate was, however, very small ($\hat{\omega}^2 = 0$). In general, it can be concluded that, for the person-fit statistics considered in this study, the proportion of aberrant response vectors in the data set did not explain the variation in the detection rates.

Factor AnsPC, however, had a large effect on the detection rates for all statistics considered (see Table 2, columns ''AnsPC''). In general, detection rates increased significantly when the sequence of aberrant item scores increased, for all person-fit statistics. This increase was larger for the CUSUM-based statistics than for the non-CUSUM-based statistics. Figures 7 and 8 display means plots which compare the effect of AnsPC on the detection rates of three CUSUM-based and the l_z statistic, for a fixed proportion of aberrant response vectors of 1%. For the random guessing and the cheating settings, the detection rates in general improved more with AnsPC for the CUSUM-based person-fit statistics than for the l_z statistic. These results illustrate that CUSUMs are more sensitive in detecting sequences of aberrant scores: Increasing the length of the sequences benefits CUSUM's approaches to detect the aberrant behavior.

Tables 3 and 4 show detection rates for each person-fit statistic under random guessing and cheating, respectively. For the van Krimpen-Stoop and Meijer's CUSUM statistics (columns 3-4 and 11-12 in Tables 3 and 4), it can be verified that C^{LR}_{VM} performed better than did the one-sided CUSUMs. Thus, C^{LR}_{VM} is a valid alternative for detection of these types of aberrancies. Comparing the CUSUM statistics from Armstrong and Shi (2009; columns 5-6 and 13-14 in Tables 3 and 4) shows that C^{L} seemed to perform better than C^{LR} under random guessing. However, C^{LR} performed better than C^{U} under cheating, especially when the length of the sequence of aberrant scores is moderately low (AnsPC = 5%, 10%). In general, Armstrong and Shi's CUSUM statistics outperformed the one- and two-sided CUSUM proposed by van

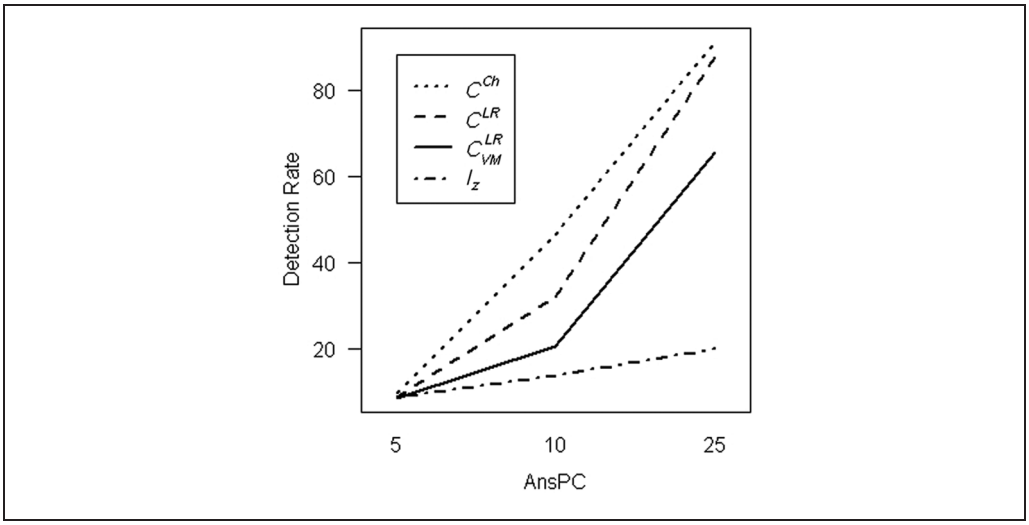

Figure 8. Cheating: Effect of the length of the sequence of aberrant item scores on detection rates for four person-fit statistics ($\alpha = .05$, SubPC = 1%)
Note: SubPC = percentage of examinees who had their response vector altered.

Krimpen-Stoop and Meijer (2000). This result is consistent with results reported in Armstrong and Shi.

The rate of false negatives can be estimated using the values in Tables 3 and 4, and using the formula (100%− detection rate). Large rates of false negatives were encountered, especially for low values of aberrance rate (indicated by AnsPC). The problem of large rates of false negatives is not exclusive to CUSUM-based statistics (see, for example, results on cheating detection reported in Belov, 2011).

Likelihood-Based CUSUMs

To assess whether the alternative models for p_i^* defined in Equations 14 and 15 affected the detection rate for the CUSUMs based on log-likelihood ratios, the detection rates of C^{RR} (for random guessing) and C^{Ch} (for cheating) were compared with the detection rates of C^L, C^U, and C^{LR}. The results are summarized in Columns 5-7 and 13-15 in Tables 3 and 4. Both C^{RR} and C^{Ch} performed better than the other CUSUM statistics. These results show how important it is to properly specify the model for the probability of correct response under aberrant behavior, p_i^*. The performance of the quadratic models defined in Equations 9 and 10 is not superior to the performance of the simpler models defined in Equations 14 and 15, respectively. In particular, the simple model that was proposed for detecting cheating resulted in high detection rates for different proportions of aberrant response vectors and for different lengths of the aberrant sequences. Thus, C^{Ch} is a simple statistic which performed quite well in terms of detection of aberrant behavior.

Statistics U, W, and l_z performed much worse when compared with the remaining statistics (Columns 8-10 and 16-18 in Tables 3 and 4). See also Figures 7 and 8 for the l_z statistic in particular. CUSUMs performed better than did statistics U, W, and l_z because the authors focused on sequences of aberrant item scores. The longer the sequences, the better the CUSUMs performed when compared with the traditional person-fit statistics. When the sequences of aberrant

Table 3. Random Guessing: Detection of Aberrance

		$\alpha = .01$								$\alpha = .05$							
SubPC	AnsPC	C^{-}	C_{VM}^{LR}	C^{L}	C^{LR}	C^{RR}	U	W	l_z	C^{-}	C_{VM}^{LR}	C^{L}	C^{LR}	C^{RR}	U	W	l_z
1	5	1.2	1.1	2.2	1.2	2.7	1.9	1.4	1.7	7.0	6.4	8.2	5.6	9.8	10.8	10.7	10.6
1	10	1.5	2.8	7.5	5.0	13.7	5.6	5.3	5.7	9.1	12.3	27.3	17.8	36.0	16.0	15.2	15.6
1	25	4.0	30.6	49.6	36.9	54.5	10.4	10.3	11.8	28.3	52.9	73.8	56.8	74.7	26.8	25.0	26.0
5	5	1.1	1.2	1.6	1.3	2.4	2.3	2.2	2.1	7.1	7.4	9.6	6.5	11.9	10.7	10.2	10.3
5	10	1.6	3.4	7.8	5.7	13.5	5.1	4.4	4.4	9.1	13.6	26.3	17.5	34.2	14.9	14.0	14.3
5	25	4.8	31.1	50.1	37.8	56.1	8.9	8.7	9.6	29.3	55.7	75.5	58.8	76.1	24.5	23.7	24.5
10	5	1.5	1.7	1.9	1.2	2.5	2.3	2.1	2.1	7.2	7.8	9.5	6.0	11.6	10.4	9.7	9.9
10	10	1.8	3.5	8.0	5.4	13.8	4.3	3.7	3.8	9.3	13.4	26.7	17.7	34.6	14.3	13.3	13.6
10	25	4.5	30.0	48.5	36.2	54.0	9.2	8.7	9.8	28.7	54.7	74.4	57.2	75.4	23.9	23.2	24.0

Note: SubPC = percentage of examinees who had their response vector altered; AnsPC = percentage of items that were altered. Values are mean percentages of aberrant score vectors which were flagged by each statistic (over the 10 replications).

Table 4. Cheating: Detection of Aberrance

		$\alpha = .01$								$\alpha = .05$							
SubPC	AnsPC	C^{-}	C_{VM}^{LR}	C^{U}	C^{Ch}	C^{RR}	U	W	l_z	C^{-}	C_{VM}^{LR}	C^{U}	C^{Ch}	C^{RR}	U	W	l_z
1	5	0.2	2.8	0.2	2.3	1.6	2.2	2.2	2.0	1.6	8.5	4.1	8.8	9.6	9.1	9.4	8.6
1	10	1.0	7.7	3.6	14.1	18.5	4.7	4.9	4.9	3.5	20.6	21.4	31.9	46.4	13.6	13.9	13.7
1	25	19.1	43.8	68.3	70.5	79.5	8.3	8.1	7.9	49.1	65.5	89.8	87.4	91.1	19.3	19.8	19.9
5	5	0.3	2.1	0.3	1.9	1.4	2.2	2.2	2.3	1.3	8.3	2.8	8.5	8.0	10.2	9.6	9.7
5	10	0.5	8.3	3.1	13.9	18.1	3.6	3.6	3.8	2.9	19.5	20.5	31.8	46.0	14.5	13.8	14.1
5	25	18.4	44.6	65.5	69.6	78.0	9.3	9.7	9.5	46.4	64.5	88.8	86.7	91.1	21.8	21.5	21.7
10	5	0.3	2.6	0.4	2.1	1.2	2.0	2.3	2.4	1.4	8.9	2.7	8.8	8.1	9.4	9.2	9.3
10	10	0.5	7.8	2.7	12.9	16.3	3.0	3.1	3.2	2.9	19.4	18.6	31.1	43.3	13.0	12.3	12.7
10	25	18.8	43.5	65.0	68.9	77.7	8.6	8.7	8.9	47.3	63.5	88.7	86.4	91.1	20.9	20.6	20.6

Note: SubPC = percentage of examinees who had their response vector altered; AnsPC = percentage of items that were altered. Values are mean percentages of aberrant score vectors which were flagged by each statistic (over the 10 replications).

scores were small (AnsPC = 5%), the statistics U, W, and l_z performed similarly or even better than the CUSUMs.

Examples of CUSUM Charts

Appendix B shows CUSUM charts for two examinees, one with normal answering behavior and the other with aberrant answering behavior, for three CUSUM statistics (C_{VM}^{LR}, C^{LR}, C^{RR}). Each chart represents the 1% and 5% control limits through horizontal lines; a response pattern is flagged as aberrant whenever the CUSUM series crosses a control limit. The section of the response vector which was altered to imitate random guessing comprised Items 25 through 50 (AnsPC = 25%). Inspecting these control charts gives a detailed picture of the performance of the examinees during the test. The charts for the C_{VM}^{LR} and C^{LR} statistics clearly show a steep increase of the statistics between Items 25 and 50, whereas the scores of the C^{RR} statistic show a pronounced decrease in the same section of the test. Note that these patterns are much different from the regular response patterns for normal examinees. The three charts give the same impression: Some odd behavior was detected between Items 25 and 50 of the test for one of the examinees. C_{VM}^{LR} and C^{LR} are two-sided statistics, hence it is more difficult to verify what type of aberrant behavior—under- or overperformance—occurred. However, the control chart for C^{RR} seems to indicate random guessing, or some other type of underperformance behavior.

Discussion

As recently discussed by Green (2011, p.173), ''A posttest search for anomalous response patterns might yield useful information for test developers.''In the present study, the authors discussed and refined different tools that can help practitioners to conduct such a search. Depending on the type of aberrant response behavior, a researcher can choose one of the methods discussed in this article. As this simulation study showed, the two-sided extension of the statistics proposed by van Krimpen-Stoop and Meijer (2000) and the newly proposed statistics for cheating and random response behavior may be good alternatives to existing procedures. Furthermore, although there are some theoretical complications in the statistics proposed by Armstrong and Shi (2009), the performance of these statistics was satisfactory in many simulated conditions.

This simulation study focused on two types of aberrant behavior—random guessing and cheating. It is interesting to observe that these types of behaviors are not necessarily complementary to each other but that they can be related. Belov (2011) showed how detected random guessing may be an indication of cheating behavior. Suppose that a test consists of an operational part (equal for all candidates) and a variable part (individually tailored) and that test takers cannot distinguish between the parts. A test taker who copies answers might display an item response sequence on the variable part which appears as random guessing, although he actually cheated. Belov also discussed scenarios where incorrect alignment or shift error behaviors might also appear as random guessing behavior. These examples illustrate that it is important to analyze all person-fit measurement results *after* the analysis has been conducted, for example, by looking at seating charts and administered items, and by possibly interviewing proctors.

Note that only moderate aberrance rates were considered in this simulation study (AnsPC = 5%, 10%, 25%). As Figures 7 and 8 show, the detection rates increased with the aberrance rate. St-Onge, Valois, Abdous, and Germain (2011) observed, however, that detection rates may decrease for high aberrance rates (larger than 40%), for some person-fit statistics. Future research is needed to indicate how CUSUMs are affected by high aberrance rates.

The authors used a completely crossed design in this study with 10 replications per cell. They did not use more replications in their study due to the extensive computation time required. They do believe that the effects reported in the current article would not change markedly for larger numbers of replications per cell.

In future studies, the usefulness of these types of statistics should be further explored. One potentially interesting application is in the field of psychological testing in personnel selection. Due to its cost-effectiveness and efficiency, a recent development in this area is the use of unproctored Internet testing. A candidate is invited to take a test at his or her place of convenience (e.g., at home), the test is administered at a computer through the internet, the candidate gets a score, and when this score is higher than some prespecified cutoff score, a candidate is invited to take a short version of the test in proctored conditions (e.g., at the office of the selection company). Guo and Drasgow (2010) suggested different statistical methods to investigate whether total test scores in both conditions are similar (and the candidate did not cheat on the first administration). In addition to these methods, the following person-fit CUSUM procedure can be conducted (see Tendeiro, Meijer, Schakel, & Maij-de Meij, in press). First, estimate the candidate's latent trait value on the unproctored test. Second, use this latent trait estimate in a CUSUM procedure together with the realized items' scores on the proctored test to investigate consistency of item answering. When a person is answering according to his latent trait value (no cheating), a normal response pattern will result. However, when a candidate answered the unproctored test with, for example, the help of someone else, an aberrant item score pattern may result at the proctored administration. Other applications may be in educational testing where preknowledge of items on parts of the test may be identified with the CUSUM procedures discussed in the present article.

Appendix A

Proof of the Invariance Property of the Quadratic Model

The authors wish to prove the invariance property, which states that the estimation of the quadratic model for p_i^* is not affected by changes in the discrimination and difficulty item parameters a_i and b_i, respectively. They sketch a proof for the upper-aberrance situation; the lower-aberrance situation is completely analogous. Please consult the estimation algorithm for the quadratic model previously presented in this article for notation.

First, the authors observe that $\theta_i^{\max}$ (the value of θ which maximizes the information function $I_i(\theta)$) and θ_i^* are defined by

$$\theta_i^{\max} = b_i + \frac{1}{a_i}\ln\left(\frac{1+\sqrt{1+8c_i}}{2}\right) \quad \text{and} \quad \theta_i^* = \theta_i^{\max} + \frac{v'}{\sqrt{I_i(\theta_i^{\max})}}, \tag{A1}$$

where

$$I_i(\theta) = a_i^2 \cdot \frac{1-p_i(\theta)}{p_i(\theta)} \cdot \left(\frac{p_i(\theta)-c_i}{1-c_i}\right)^2, \tag{A2}$$

and p_i is the 3PL model defined in Equation 8. Evaluating the value of p_i when $\theta = \theta_i^{\max}$ and $\theta = \theta_i^*$ yields expressions which are independent from a_i and b_i:

$$p_i(\theta_i^{\max}) = \frac{1+2c_i+\sqrt{1+8c_i}}{3+\sqrt{1+8c_i}} \tag{A3}$$

and

$$p_i(\theta_i^*) = c_i + (1 - c_i) \Big/ \left[1 + \left(\frac{2}{1 + \sqrt{1 + 8c_i}} \right) \cdot \mathrm{Exp} \left(\frac{-v'}{2\sqrt{\frac{(1-c_i)(1 + 4c_i + \sqrt{1+8c_i})}{(3+\sqrt{1+8c_i})^2 (1 + 2c_i + \sqrt{1+8c_i})}}} \right) \right]. \quad \text{(A4)}$$

Hence, $\lambda_i = \frac{1-p_i(\theta_i^*)}{1-p_i(\theta_i^{\max})}$ is also independent from a_i and b_i. Finally, the authors observe that estimating coefficients r_i^U, s_i^U, and t_i^U is done by solving the equations $g_i^U(1) = 1$, $g_i^U(p_i(\theta_i^{\max})) = p_i(\theta_i^*)$, and $g_i^U(c_i) = (1 - v'')c_i + v''(p_i(\theta_i^*) + \lambda_i(c_i - p_i(\theta_i^{\max})))$ with respect to the three unknowns r_i^U, s_i^U, and t_i^U. All equations in the system depend only on c_i, v', and v''. This finishes the proof.

Appendix B

CUSUM Control Charts: Random Responding

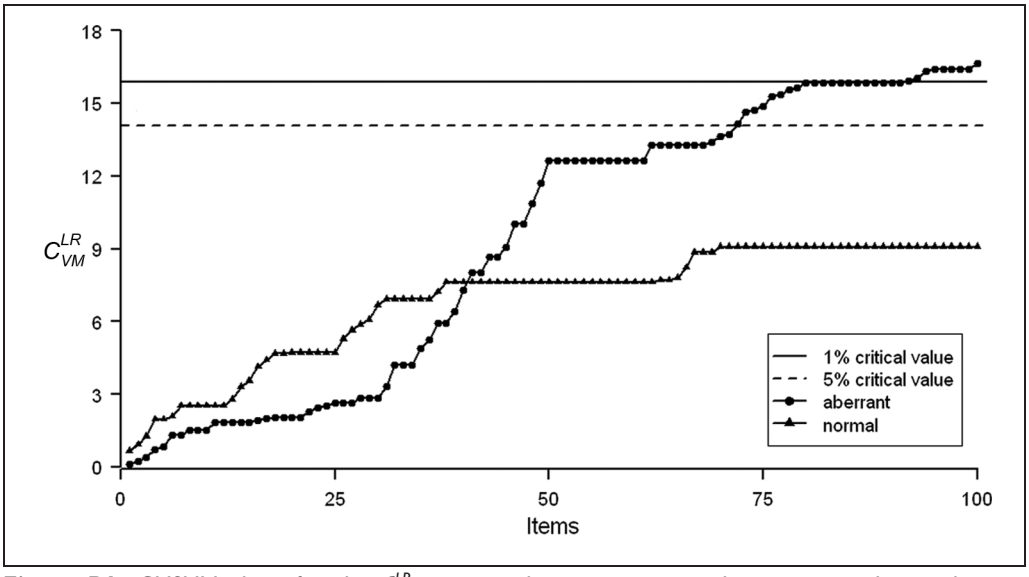

Figure B1. CUSUM chart for the C_{VM}^{LR} statistic showing one normal examinee and one aberrant examinee ("random guessing" inputted on Items 25 to 50)
Note: CUSUM = CUmulative SUM. Example from simulation study.

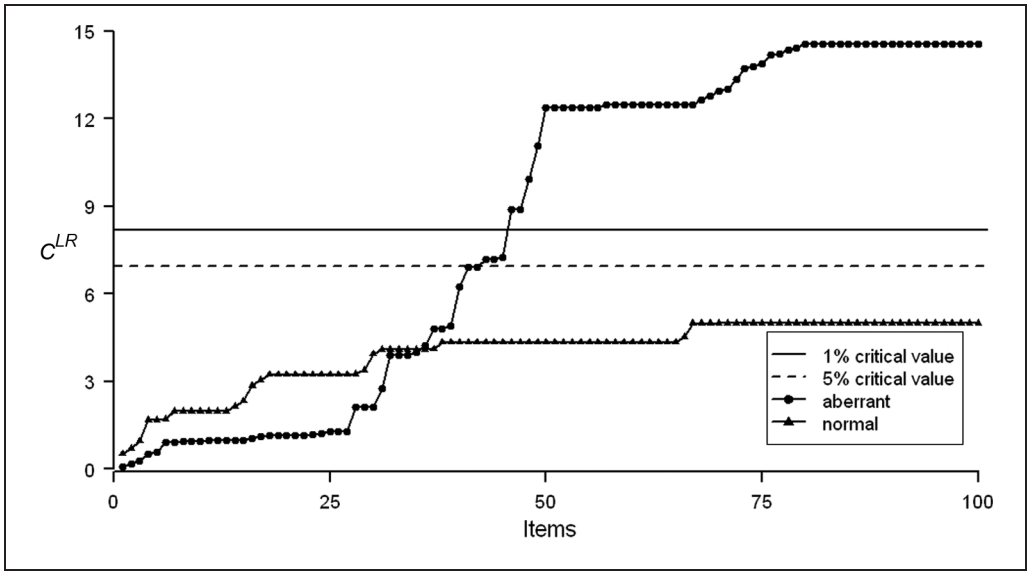

Figure B2. CUSUM chart for the C^{LR} statistic showing one normal examinee and one aberrant examinee ("random guessing" inputted on Items 25 to 50)
Note: CUSUM = CUmulative SUM. Example from simulation study.

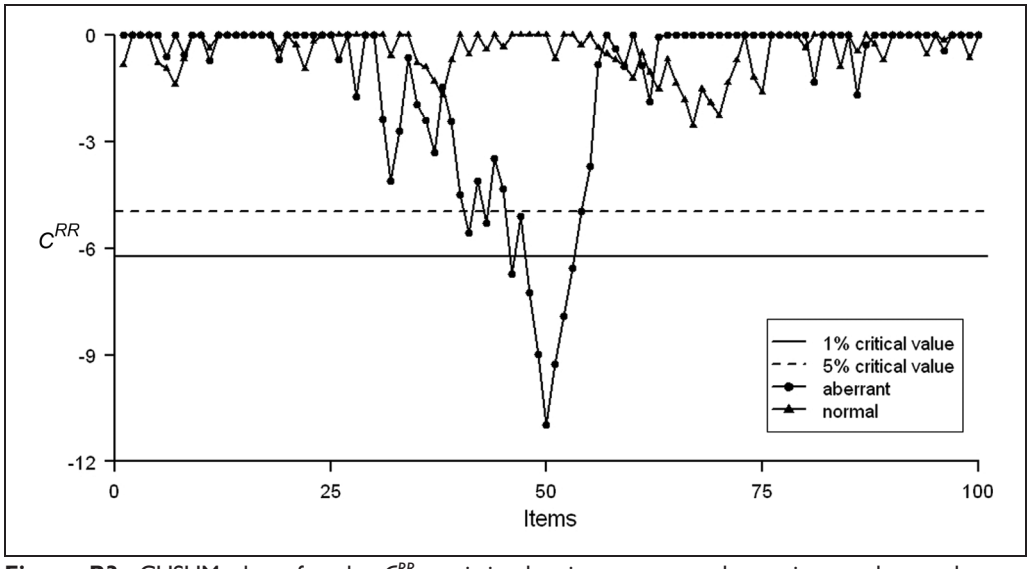

Figure B3. CUSUM chart for the C^{RR} statistic showing one normal examinee and one aberrant examinee ("random guessing" inputted on Items 25 to 50)
Note: CUSUM = CUmulative SUM. Example from simulation study.

Acknowledgment

The authors thank two anonymous reviewers and the editor for their help with improving this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person-fit. *Applied Psychological Measurement, 33*, 391-410.

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement, 35*, 495-517.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910-919.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47-64.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Green, B. F. (2011). A comment on early student blunders on computer-based adaptive tests. *Applied Psychological Measurement, 35*, 165-174.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18*, 351-364.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*, 843-877.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3-8.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment, 90*, 227-238.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2010). Detecting person misfit in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 315-329). New York, NY: Springer.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.

Page, E. S. (1954). Continuous inspection procedures. *Biometrika, 41*, 100-115.

R Development Core Team. (2009). *R: A language and environment for statistical computing* (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*, 83-101.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement, 35*, 419-432.

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (in press). Using cumulative sum statistics to detect inconsistencies in unproctored Internet testing. *Educational and Psychological Measurement*.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detection of person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston, MA: Kluwer-Nijhoff.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

Wright, B. D. (1980). Afterward. In G. Rasch (Ed.), *Probabilistic models for some intelligence and attainment tests: With foreword and afterword by Benjamin D. Wright*. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: Mesa Press.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG for Windows: Multiple-group IRT analysis and test maintenance for binary items (Version 3.0) [Computer program]. Chicago, IL: Scientific Software International.